W53

An apparatus for reallocating logical to physical disk devices using a storage controller and method of the same

Patent Number: EP0801344, A3

Publication date: 1997-10-15

Inventor(s): YAMAMOTO YASUTOMO (JP); SATOH TAKAO (JP); YAMAMOTO AKIRA (JP)

Applicant(s):: HITACHI LTD (JP)

Requested Patent: JP9274544

Application Number: EP19970105448 19970402

Priority Number(s): JP19960085370 19960408

IPC Classification: G06F3/06

EC Classification: G06F3/06M, G06F11/10M, G06F11/20L

Equivalents: US5956750

Abstract

A storage controller (104) includes that it calculates an access frequency (500) of each logical disk (200); that it selects first logical disk device of which the access frequency exceeds a first predetermined value, the first logical disk device being allocated to a first physical disk device; that it selects a second logical disk device which has the access frequency equal to or less than a second predetermined value, the second logical disk device being allocated to a second physical disk device; and that it reallocates the first and second logical devices to the second and the first physical disk devices, respectively.

(19)日本国特許庁（ＪＰ）　(12)公開特許公報(A)　(11)特許出願公開番号

特開平9－274544

(43)公開日　平成９年(1997)10月21日

(51)Int.Cl.*		識別記号	庁内整理番号	ＦＩ		技術表示箇所
Ｇ０６Ｆ	3/06	５４０		Ｇ０６Ｆ	3/06	５４０
		３０１				３０１Ｘ
		３０２				３０２Ｅ
		３０４				３０４Ｎ
	12/08	３２０	7623－5Ｂ		12/08	３２０

審査請求　未請求　請求項の数９　ＯＬ　（全 14 頁）

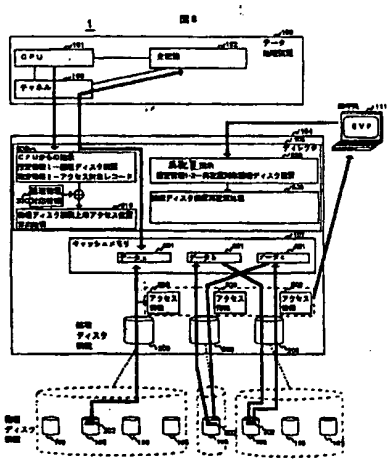
(21)出願番号　特願平8－85370

(22)出願日　平成８年(1996)４月８日

(71)出願人　000005108
株式会社日立製作所
東京都千代田区神田駿河台四丁目６番地
(72)発明者　山本　康友
神奈川県川崎市麻生区王禅寺1099番地　株式会社日立製作所システム開発研究所内
(72)発明者　山本　彰
神奈川県川崎市麻生区王禅寺1099番地　株式会社日立製作所システム開発研究所内
(72)発明者　佐藤　孝夫
神奈川県小田原市国府津2880番地　株式会社日立製作所ストレージシステム事業部内
(74)代理人　弁理士　有近　紳志郎

(54)【発明の名称】　記憶制御装置

(57)【要約】
【課題】　アクセス性能を向上する。
【解決手段】　各論理ディスク装置２００に対するアクセス情報５００を採取し、そのアクセス情報５００をＳＶＰ１１１を通じて保守員に提示する。保守員の再配置指示６２０があると、指示された２つの論理ディスク装置２００の間で物理ディスク装置１０５を配置替えし、全データを連続的に格納し直す。
【効果】　アクセス頻度の高い論理ディスク装置をより高速な物理ディスク装置へ再配置することが出来る。シーケンシャルアクセスの比率の高い論理ディスク装置をよりシーケンシャルアクセス性能の高い物理ディスク装置へ再配置することが出来る。



【特許請求の範囲】
【請求項１】　データ処理装置が直接アクセスを行う論理的記憶装置を実際にデータを記憶する物理的記憶装置に配置し、前記データ処理装置と前記物理的記憶装置の間のデータ転送を制御する記憶制御装置において、予め定めた指標に基づいて前記論理的記憶装置を前記物理的記憶装置に再配置すると共に再配置先の物理的記憶装置にデータを連続的に格納する論理的記憶装置再配置手段を有することを特徴とする記憶制御装置。
【請求項２】　データ処理装置が直接アクセスを行う論理的記憶装置と実際にデータを記憶する物理的記憶装置とを対応付け、前記データ処理装置と前記物理的記憶装置の間のデータ転送を制御する記憶制御装置において、前記データ転送の制御の運用中にデータ処理装置の論理的記憶装置へのアクセス情報を指標として採取するアクセス情報採取手段と、前記指標に基づいて前記論理的記憶装置を前記物理的記憶装置に再配置すると共に再配置先の物理的記憶装置にデータを連続的に格納する論理的記憶装置再配置手段とを有することを特徴とする記憶制御装置。
【請求項３】　請求項２に記載の記憶制御装置において、前記アクセス情報が、前記データ処理装置から前記論理的記憶装置へのアクセス頻度情報を含むことを特徴とする記憶制御装置。
【請求項４】　請求項２または請求項３に記載の記憶制御装置において、前記アクセス情報が、前記データ処理装置から前記論理的記憶装置へのアクセスパターン情報を含むことを特徴とする記憶制御装置。
【請求項５】　請求項１に記載の記憶制御装置において、前記指標が、前記論理的記憶装置に求められる信頼性であることを特徴とする記憶制御装置。
【請求項６】　請求項１から請求項５のいずれかに記載の記憶制御装置において、前記指標を保守員に提示する指標提示手段と、保守員からの再配置指示を受け付ける再配置指示受付手段とを具備したことを特徴とする記憶制御装置。
【請求項７】　請求項１から請求項５のいずれかに記載の記憶制御装置において、データ処理装置からの再配置指示を受け付ける再配置指示受付手段を具備したことを特徴とする記憶制御装置。
【請求項８】　請求項１から請求項５のいずれかに記載の記憶制御装置において、前記指標に基づいて再配置の要否を決定する再配置要否決定手段を具備したことを特徴とする記憶制御装置。
【請求項９】　請求項１から請求項８のいずれかに記載の記憶制御装置において、再配置中の論理的記憶装置にデータ処理装置からのアクセスがあったとき、再配置中の論理的記憶装置の再配置完了領域と再配置未完領域とを識別し、前記アクセス位置が前記再配置完了領域ならば再配置先の論理的記憶装置にアクセスさせ、前記アクセス位置が前記再配置未完領域ならば当該論理的記憶装置にアクセスさせるアクセス位置切替手段をさらに具備したことを特徴とする記憶制御装置。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】本発明は、記憶制御装置に関し、さらに詳しくは、シーケンシャルアクセスの場合やランダムアクセルでヒット率が低い場合でもアクセス性能を向上することが出来る記憶制御装置およびデータの信頼性を向上することが出来る記憶制御装置に関する。特に、本発明は、ディスクアレイ向きの高機能ディスク装置、その高機能ディスク装置とディスク制御装置とにより構成される記憶装置サブシステム、およびその記憶装置サブシステムとデータ処理装置とにより構成される情報処理システムに有用である。
【０００２】
【従来の技術】シカゴのイリノイ大学で開かれた「ACM SIGMOD」会議において発表された論文「D. Patterson, G. gibson, and R. H. Kartz:A Case for Redundant Arrays of Inexpensive Disks (RAID), ACM SIGMOD Conference, Chicago, IL, (June 1988), pp. 109-116」は、ディスクアレイ上のデータ配置に関する技術を開示している。
【０００３】また、特開平７－８４７３２号公報では、ディスク装置の一部をディスクキャッシュの如く用いる技術が開示されている。具体的には、ディスク装置を一時的にデータを格納するテンポラリ領域と最終的にデータを書き込む領域とに分け、更新データはパリティを生成せずに一旦テンポラリ領域に二重書きし、非同期にパリティ生成し、最終領域に書き込む。
【０００４】一方、電気情報通信学会技術研究報告「ＤＥ９５－６８（茂木他：Hot Mirroring を用いたディスクアレイのディスク故障時の性能評価、１９９５年１２月、電気情報通信学会技報 Vol. 95-No. 407, pp. 19-24）」には、アクセス頻度の違いにより、データを保持するＲＡＩＤレベルを動的に変更する技術が開示されている。具体的には、ディスク装置をＲＡＩＤ１構成の部分とＲＡＩＤ５構成の部分に分け、ライトアクセスのあったデータを優先的にＲＡＩＤ１構成の部分に格納するようにデータの格納位置を動的に変更することにより、アクセス頻度の高いデータはＲＡＩＤ１構成の部分に格納し、アクセス頻度の低いものはＲＡＩＤ５構成の部分に格納するように出来る。この技術によれば、記憶容量の異なる物理ディスク装置やＲＡＩＤレベルの異なる物理ディスク装置を記憶装置サブシステム内で混在させることが可能であり、論理ディスク装置内のデータを、そのアクセス頻度やアクセスパターンなどの指標に基づいて、任意の物理ディスク装置に格納することが出来る。また、アクセス頻度の高いデータを、より高速な物理ディスク装置に格納するように、動的に格納位置を変更することも出来る。なお、ＲＡＩＤ１のディスクアレイ

1143

は、データ処理装置からの書き込みデータに対して、その複製をミラーと呼ばれる副ディスク装置に書き込み、データの信頼性を確保する。冗長データが元のデータの複製であるため、冗長データ作成のオーバヘッドが小さく、アクセス性能が良い。但し、物理的記憶装置の使用効率は、５０％と低い。一方、ＲＡＩＤ５のディスクアレイは、データ処理装置からの複数の書き込みデータに対して、パリティと呼ばれる冗長データを作成する。パリティ作成時に更新前データと更新前パリティのリードが必要なため、冗長データ作成のオーバヘッドが大きく、アクセス性能は悪い。但し、複数のデータに対して１つのパリティを作成するため、記憶装置の使用効率はＲＡＩＤ１に比べ高い。

【０００５】

【発明が解決しようとする課題】上記従来技術では、アクセスするデータ単位でデータの格納位置の変更を行うため、データ処理装置が直接アクセスを行う論理ディスク装置上では連続なデータが、実際にデータを記憶する物理ディスク装置上では非連続となってしまう。このため、一連のデータをリード／ライトするシーケンシャルアクセスの場合、実際には複数データをまとめてリード／ライトできなくなり、アクセス性能の低下を招く問題点がある。

【０００６】一方、上記報告「ＤＥ９５－６８」の従来技術では、ライトの度に、アクセス頻度が低いと判断したデータをＲＡＩＤ１構成の部分からＲＡＩＤ５構成の部分に移し、空いたＲＡＩＤ１構成の部分にライトデータを書き込むため、アクセスパターンがランダムアクセルでヒット率が低い場合には、ＲＡＩＤ１構成の部分に移したデータの多くは再びＲＡＩＤ５構成の部分に戻されることになる。このため、ヒット率が低い場合、アクセス性能の向上は期待できず、逆にデータを移す処理のオーバヘッドがアクセス性能の低下を引き起こす問題点がある。

【０００７】また、上記の従来技術では、データの信頼性の向上については全く考慮されていない問題点がある。

【０００８】そこで、本発明の第１の目的は、シーケンシャルアクセスの場合やランダムアクセルでヒット率が低い場合でも、アクセス性能を向上することが出来る記憶制御装置を提供することにある。また、本発明の第２の目的は、データの信頼性を向上することが出来る記憶制御装置を提供することにある。

【０００９】

【課題を解決するための手段】第１の観点では、本発明は、データ処理装置が直接アクセスを行う論理的記憶装置を実際にデータを記憶する物理的記憶装置に配置し、前記データ処理装置と前記物理的記憶装置の間のデータ転送を制御する記憶制御装置において、予め定めた指標に基づいて前記論理的記憶装置を前記物理的記憶装置に再配置すると共に再配置先の物理的記憶装置にデータを連続的に格納する論理的記憶装置再配置手段を有することを特徴とする記憶制御装置を提供する。上記第１の観点による記憶制御装置では、アクセスするデータ単位でデータの格納位置の変更を行うのではなく、論理的記憶装置を単位として物理的記憶装置への再配置を行い、且つ、再配置先の物理的記憶装置にデータを連続的に格納する。従って、シーケンシャルアクセスの場合でも、アクセス性能を向上することが出来る。また、ライトの度にデータの格納位置の変更を行うのではなく、予め定めた指標に基づいて前記再配置を行うから、ランダムアクセルでヒット率が低い場合でも、アクセス性能を向上することが出来る。

【００１０】第２の観点では、本発明は、データ処理装置が直接アクセスを行う論理的記憶装置と実際にデータを記憶する物理的記憶装置とを対応付け、前記データ処理装置と前記物理的記憶装置の間のデータ転送を制御する記憶制御装置において、前記データ転送の制御の運用中にデータ処理装置の論理的記憶装置へのアクセス情報を指標として採取するアクセス情報採取手段と、前記指標に基づいて前記論理的記憶装置を前記物理的記憶装置に再配置すると共に再配置先の物理的記憶装置にデータを連続的に格納する論理的記憶装置再配置手段とを有することを特徴とする記憶制御装置を提供する。上記第２の観点による記憶制御装置では、アクセスするデータ単位でデータの格納位置の変更を行うのではなく、論理的記憶装置を単位として物理的記憶装置への再配置を行い、且つ、再配置先の物理的記憶装置にデータを連続的に格納する。従って、シーケンシャルアクセスの場合でも、アクセス性能を向上することが出来る。また、ライトの度にデータの格納位置の変更を行うのではなく、アクセス情報を採取し、それを統計的に利用して前記再配置を行うから、ランダムアクセルでヒット率が低い場合でも、アクセス性能を向上することが出来る。

【００１１】第３の観点では、本発明は、上記構成の記憶制御装置において、前記アクセス情報が、前記データ処理装置から前記論理的記憶装置へのアクセス頻度情報を含むことを特徴とする記憶制御装置を提供する。上記第３の観点による記憶制御装置では、アクセス頻度の高い論理的記憶装置をより高速な物理的記憶装置へ再配置することが出来る。従って、アクセス性能を向上することが出来る。

【００１２】第４の観点では、本発明は、上記構成の記憶制御装置において、前記アクセス情報が、前記データ処理装置から前記論理的記憶装置へのアクセスパターン情報を含むことを特徴とする記憶制御装置を提供する。上記第４の観点による記憶制御装置では、シーケンシャルアクセスの比率の高い論理的記憶装置をよりシーケンシャルアクセス性能の高い物理的記憶装置へ再配置することが出来る。従って、アクセス性能を向上することが出来る。

【００１３】第５の観点では、本発明は、上記構成の記憶制御装置において、前記指標が、前記論理的記憶装置に求められる信頼性であることを特徴とする記憶制御装置を提供する。上記第５の観点による記憶制御装置では、信頼性が高いことが求められる論理的記憶装置をより信頼性の高い物理的記憶装置へ再配置することが出来る。従って、データの信頼性を向上することが出来る。

【００１４】第６の観点では、本発明は、上記構成の記憶制御装置において、前記指標を保守員に提示する指標提示手段と、保守員からの再配置指示を受け付ける再配置指示受付手段とを具備したことを特徴とする記憶制御装置を提供する。上記第６の観点による記憶制御装置では、保守員が再配置指示を入力できるため、非常に柔軟に前記再配置を行うことが出来る。

【００１５】第７の観点では、本発明は、上記構成の記憶制御装置において、データ処理装置からの再配置指示を受け付ける再配置指示受付手段を具備したことを特徴とする記憶制御装置を提供する。上記第７の観点による記憶制御装置では、データ処理装置が再配置指示を入力できるため、保守員では判断不可能な高度の条件下で前記再配置を行うことが出来る。

【００１６】第８の観点では、本発明は、上記構成の記憶制御装置において、前記指標に基づいて再配置の要否を決定する再配置要否決定手段を具備したことを特徴とする記憶制御装置を提供する。上記第８の観点による記憶制御装置では、記憶制御装置が再配置指示を自己決定するため、保守員やデータ処理装置に負担をかけなくて済む。

【００１７】第９の観点では、本発明は、上記構成の記憶制御装置において、再配置中の論理的記憶装置にデータ処理装置からのアクセスがあったとき、再配置中の論理的記憶装置の再配置完了領域と再配置未完領域とを識別し、前記アクセス位置が前記再配置完了領域ならば再配置先の論理的記憶装置にアクセスさせ、前記アクセス位置が前記再配置未完領域ならば当該論理的記憶装置にアクセスさせるアクセス位置切替手段を具備したことを特徴とする記憶制御装置を提供する。上記第９の観点による記憶制御装置では、再配置中の論理的記憶装置の再配置完了領域と再配置未完領域とを識別し、データ処理装置からのアクセス位置を切り替えるから、データ処理装置と物理的記憶装置の間のデータ転送を運用中に再配置を行うことが出来る。

【００１８】

【発明の実施の形態】以下、本発明の実施形態を説明する。なお、これにより本発明が限定されるものではない。

【００１９】－第１の実施形態－

第１の実施形態は、各論理ディスク装置のアクセス情報を記憶制御装置で採取し、ＳＶＰ（サービスプロセッサ）を通じて保守員に提示し、このアクセス情報に基づく保守員の再配置指示により、論理ディスク装置の物理ディスク装置への再配置を行うものである。

【００２０】図１は、本発明の第１の実施形態にかかる記憶制御装置を含む情報処理システムのブロック図である。この情報処理システム１は、データ処理装置１００と、記憶制御装置１０４と、１台以上の物理ディスク装置１０５と、ＳＶＰ１１１とを接続してなっている。

【００２１】前記データ処理装置１００は、ＣＰＵ１０１と、主記憶１０２と、チャネル１０３とを有している。

【００２２】前記記憶制御装置１０４は、１つ以上のディレクタ１０６と、キャッシュメモリ１０７と、ディレクトリ１０８と、不揮発性メモリ１０９と、不揮発性メモリ管理情報１１０と、論理物理対応情報３００と、論理ディスク装置情報４００と、アクセス情報５００を有している。前記ディレクタ１０６は、データ処理装置１００のチャネル１０３と物理ディスク装置１０５の間のデータ転送、データ処理装置１００のチャネル１０３と前記キャッシュメモリ１０７の間のデータ転送および前記キャッシュメモリ１０７と物理ディスク装置１０５の間のデータ転送を行う。前記キャッシュメモリ１０７には、物理ディスク装置１０５の中のアクセス頻度の高いデータをロードしておく。このロード処理は、前記ディレクタ１０６が実行する。ロードするデータの具体例は、データ処理装置１００のＣＰＵ１０１のアクセス対象データや、このアクセス対象データと物理ディスク装置１０５上の格納位置が近いデータ等である。前記ディレクトリ１０８は、前記キャッシュメモリ１０７の管理情報を格納する。前記不揮発性メモリ１０９は、前記キャッシュメモリ１０７と同様に、物理ディスク装置１０５の中のアクセス頻度の高いデータをロードしておく。前記不揮発性メモリ管理情報１１０は、前記不揮発性メモリ１０９の管理情報を格納する。前記論理物理対応情報３００は、各論理ディスク装置（図２の２００）が配置されている物理ディスク装置１０５上の位置および各物理ディスク装置１０５に配置されている論理ディスク装置（図２の２００）を示す情報である。この情報を用いて、データ処理装置１００のＣＰＵ１０１のアクセス対象データの物理ディスク装置１０５上の格納領域の算出などを行う。前記論理ディスク装置情報４００は、各論理ディスク装置（図２の２００）のアクセス可否等の状態を示す。前記アクセス情報５００は、各論理ディスク装置（図２の２００）のアクセス頻度やアクセスパターンなどの情報である。

【００２３】論理物理対応情報３００と論理ディスク情報４００は、電源断などによる消失を防ぐために不揮発の媒体に記録する。

【００２４】前記物理ディスク装置１０５は、データを記録する媒体と、記録されたデータを読み書きする装置

とから構成される。

【0025】前記SVP111は、アクセス情報500の保守員への提示や，保守員からの再配置指示620の入力の受け付けを行う。また、保守員からの情報処理システム1への指示の発信や，情報処理システム1の障害状態等の保守員への提示を行う。

【0026】図2は、論理ディスク装置200と物理ディスク装置105の関連を表わした図である。論理ディスク装置200は、データ処理装置100のCPU101が直接アクセスする見掛け上のディスク装置で、アクセス対象データが実際に格納される物理ディスク装置105と対応している。論理ディスク装置200上のデータは、シーケンシャルアクセスを考慮して、物理ディスク装置105上に連続的に配置されている。論理ディスク装置200のデータが配置されている物理ディスク装置105がディスクアレイ構成の場合、該論理ディスク装置200は複数の物理ディスク装置105と対応する。また、物理ディスク装置105の容量が論理ディスク装置200より大きく、複数の論理ディスク装置のデータを1台の物理ディスク装置105に格納できる場合には、該物理ディスク装置105は複数の論理ディスク装置200と対応する。この論理ディスク装置200と物理ディスク装置105の対応は前記論理物理対応情報300で管理される。例えば、データ処理装置100のCPU101が論理ディスク装置200のデータ201をリードする時、記憶制御装置104で論理物理対応情報300に基づき論理ディスク装置200に対応する物理ディスク装置105を求め、その物理ディスク装置105の領域内のデータ格納位置202を求め、データ転送を行う。

【0027】図3は、論理物理対応情報300を表わした図である。論理物理対応情報300は、論理ディスク構成情報310と、物理ディスク構成情報320とから構成される。前記論理ディスク構成情報310は、各論理ディスク装置200が配置されている物理ディスク装置105上の領域に関する情報であり、論理ディスク装置200から対応する物理ディスク装置105を求める時に用いる。一方、前記物理ディスク構成情報320は、各物理ディスク装置105に配置されている論理ディスク装置200に関する情報で、物理ディスク装置105から対応する論理ディスク装置200を求める時に用いる。

【0028】前記論理ディスク構成情報310は、物理ディスク装置グループ311，RAID構成312および開始位置313の組を、論理ディスク装置200の数だけ有している。前記物理ディスク装置グループ311は、当該論理ディスク装置200が配置されている物理ディスク装置105を示す情報である。前記RAID構成312は、前記物理ディスク装置グループ311のRAIDレベルを示す。前記開始位置313は、当該論理ディスク装置200が物理ディスク装置105上で配置されている先頭位置を示す。

【0029】前記物理ディスク構成情報320は、論理ディスク装置グループ321を、物理ディスク装置105の数だけ有している。前記論理ディスク装置グループ321は、当該物理ディスク装置105に配置されている論理ディスク装置200を示す。

【0030】図4は、論理ディスク情報400を表わした図である。論理ディスク情報400は、論理ディスク状態401と再配置完了ポインタ402とを、論理ディスク装置200の数だけ有している。前記論理ディスク状態401は、「正常」「閉塞」「フォーマット中」「再配置中」などの論理ディスク装置200の状態を表わす。前記再配置完了ポインタ402は、前記論理ディスク状態401が「再配置中」の時のみ有効な情報で、当該論理ディスク装置200の再配置処理を完了している領域の次の位置すなわち当該論理ディスク装置200が未だ再配置処理を終えていない領域の先頭位置を示す。「再配置中」におけるデータアクセス時、再配置完了ポインタ402よりも前の領域へのアクセスの場合には、再配置後の物理ディスク装置105へアクセスしなければならない。一方、再配置完了ポインタ402以後の領域へのアクセスの場合には、再配置前の物理ディスク装置105へアクセスしなければならない。

【0031】図5は、アクセス情報500を表わしている。アクセス情報500は、アクセス頻度情報501とアクセスパターン情報502とを、論理ディスク装置200の数だけ有している。このアクセス情報500は、記憶制御装置104，データ処理装置100，SVP111のいずれからも参照することが出来る。前記アクセス頻度情報501は、単位時間あたりの当該論理ディスク装置200へのアクセス回数を管理する。このアクセス頻度情報501は、各論理ディスク装置200の中でアクセス頻度の高いもの又は低いものを求める指標として用いる。前記アクセスパターン情報502は、当該論理ディスク装置200へのシーケンシャルアクセスとランダムアクセスの割合を管理する。このアクセスパターン情報502は、シーケンシャルアクセスが多く、よりシーケンシャル性能の高い物理ディスク装置105に再配置するのが望ましい論理ディスク装置200を求める指標として用いる。

【0032】次に、記憶制御装置104の動作を説明する。図6は、記憶制御装置104の動作を詳細に表わした図である。まず、リード／ライト処理時の動作について説明する。ディレクタ106は、通常リード／ライト処理を実行する際、CPU101からチャネル103を経由してCPUからの指示600を受け取る。このCPUからの指示600は、リード（またはライト）対象のレコードが記憶されている論理ディスク装置200を指定する指定情報1と、リード（またはライト）対象のレコードが記憶されている論理ディスク装置200内の位置（トラック，セクタ，レコード）を指定する指定情報2とを含んでいる。ディレクタ106は、物理ディスク装置上のアクセス位置算出処理（610）で、前記CPUからの指示600と論理物理対応情報300とを用いて、物理ディスク装置105上でのアクセス位置を算出する。この物理ディスク装置アクセス位置算出処理（610）については図8を参照して後で詳述する。その後、たとえばリード処理では、算出した物理ディスク装置105上のデータ格納位置202のデータをキャッシュメモリ107上に読み上げてデータ201とし、その読み上げたデータ201をチャネル103を通じて主記憶102に転送する。

【0033】次に、アクセス情報500の採取処理について説明する。CPU101からのリード／ライト処理のアクセス時に、ディレクタ106は、アクセス対象論理ディスク装置200のアクセス情報500を更新する。アクセス頻度情報501の採取は、例えば、アクセスの度に内部カウンタをカウントアップしていき、一定時間または一定回数のアクセス経過後のアクセス時に、前記内部ウンタからアクセス頻度を判定する。アクセスパターン情報502の採取は、例えば、アクセスの度に内部カウンタにシーケンシャルアクセス回数をカウントアップしていき、一定時間または一定回数のアクセス経過後のアクセス時に、前記内部カウンタからアクセスパターンを判定する。

【0034】次に、再配置指示620を説明する。保守員は、SVP111を通じて提示されたアクセス情報500を参照して、各論理ディスク装置200の再配置の必要性を検討する。この検討の結果、再配置を決定した論理ディスク装置200があれば、SVP111を通じて記憶制御装置104に対して再配置指示620を出す。この再配置指示620は、再配置対象の論理ディスク装置200を2つ指定する指示情報1－2からなる。保守員が行う検討の内容は、後述する第3の実施形態で図10を参照して説明する論理ディスク装置再配置要否決定処理（910）と同様である。

【0035】次に、論理ディスク装置再配置処理（630）を説明する。ディレクタ106は、前記再配置指示620を受けて、指定された2つの論理ディスク装置200の間で論理ディスク装置再配置処理（630）を行う。図7は、論理ディスク装置再配置処理部630の処理フロー図である。ステップ700では、論理ディスク情報400のうちの指定された2つの論理ディスク装置200の論理ディスク状態401を「再配置中」に設定する。ステップ701では、論理ディスク情報400のうちの指定された2つの論理ディスク装置200の再配置完了ポインタ402を各論理ディスク装置200の先頭位置に初期化する。ステップ702では、論理ディスク情報400のうちの指定された2つの論理ディスク装置200の再配置完了ポインタ402をチェックし、全領域の再配置が完了していなければステップ703へ進み、完了していればステップ707へ進む。

【0036】ステップ703では、再配置完了ポインタ402が示すデータ位置から再配置処理の1回の処理単位分のデータに対して物理ディスク装置105からキャッシュメモリ107上へのデータ転送を行う。ここで、1回の処理単位分のデータ量は、再配置対象の2つの論理ディスク装置200の冗長データ1つに対応する各データ量の最小公倍数に決定される。たとえば、再配置をRAID5の論理ディスク装置200とRAID1の論理ディスク装置200の間で行うならば、RAID1の論理ディスク装置200の冗長データ1つに対応するデータ量は“1”であるから、1回の処理単位分のデータ量は、RAID5の論理ディスク装置200の冗長データ1つに対応するデータ量すなわちパリティ1つに対応するデータ量に決定される。

【0037】ステップ704では、再配置対象の各論理ディスク装置200の再配置先論理ディスク装置200がパリティを有するRAIDレベルのものである場合、キャッシュメモリ107上の再配置対象の1回の処理単位分のデータ201に対してパリティを生成する。ステップ705では、キャッシュメモリ107上の再配置対象の1回の処理単位分のデータ201および前記ステップ704で作成したパリティを、再配置先の物理ディスク装置105へ書き込む。ステップ706では、1回の処理単位分だけ再配置完了ポインタ402を進める。そして、前記ステップ702に戻る。

【0038】なお、上記ステップ703，704において、データおよびパリティは、不揮発性メモリ109にも転送して二重化し、キャッシュ障害によるデータ消失を防ぐ。この理由は、上記ステップ705での書き込み時に、例えば、第1の論理ディスク装置200と第2の論理ディスク装置200のデータのうち、第1の論理ディスク装置200のデータを物理ディスク装置105（元は第2の論理ディスク装置200に配置されていた物理ディスク装置105）へ書き込んだ段階で障害によりキャッシュメモリ107上のデータがアクセス不能になったとすると、書き込みが終了してない第2の論理ディスク装置200のデータが消失するからである（元は第2の論理ディスク装置200に配置されていた物理ディスク装置105には、上記のように第1の論理ディスク装置200のデータが上書きされてしまっている）。

【0039】ステップ707では、論理物理対応情報300を更新する。すなわち、論理ディスク構成情報310と物理ディスク構成情報321を変更する。ステップ708では、論理ディスク情報400の論理ディスク状態401を元の状態に戻し、再配置処理（630）を終了する。

【0040】次に、物理ディスク装置アクセス位置算出

処理（６１０）を説明する。図8は、物理ディスク装置アクセス位置算出処理部６１０の処理フロー図である。ステップ８００では、論理ディスク情報４００のうちのアクセス対象論理ディスク装置２００の論理ディスク状態４０１が「再配置中」であるか否かをチェックし、「再配置中」ならばステップ８０１に進み、「再配置中で」なければステップ８０３に進む。

【００４１】ステップ８０１では、論理ディスク情報４００のうちのアクセス対象論理ディスク装置２００の再配置完了ポインタ４０２とアクセスデータ位置とを比較し、アクセスデータ位置が再配置完了ポインタ４０２の指す位置以後ならばステップ８０２に進み、アクセスデータ位置が再配置完了ポインタ４０２の指す位置より前ならばステップ８０３に進む。

【００４２】ステップ８０２では、当該論理ディスク装置２００の再配置先の論理ディスク装置２００をアクセス対象にする。そして、ステップ８０４へ進む。

【００４３】ステップ８０３では、当該論理ディスク装置２００をアクセス対象とする。

【００４４】ステップ８０４では、アクセス対象の論理ディスク装置２００に対応した物理ディスク装置１０５上でのアクセス位置を、論理物理対応情報３００を用いて算出する。

【００４５】以上の第１の実施形態にかかる情報処理システム１および記憶制御装置１０４によれば、アクセス情報５００に基づく保守員の判断により、アクセス頻度の高い論理ディスク装置をより高速な物理ディスク装置へ再配置することが出来る。また、シーケンシャルアクセスの比率の高い論理ディスク装置をよりシーケンシャルアクセス性能の高い物理ディスク装置へ再配置することが出来る。従って、アクセス性能を向上することが出来る。

【００４６】－第２の実施形態－

上記第１の実施形態を変形して、記憶制御装置１０４からアクセス情報５００をデータ処理装置１００に提示し、データ処理装置１００が再配置要否を決定し記憶制御装置１０４に再配置指示（６２０相当）を出すようにしてもよい。

【００４７】－第３の実施形態－

第３の実施形態は、再配置指示をＳＶＰ１１１やデータ処理装置１００から受けるのではなく、記憶制御装置１０４が自己決定するものである。

【００４８】図９は、記憶制御装置１０４の動作を詳細に表わした図である。第１の実施形態（図６）との違いは、論理ディスク再配置要否決定処理部９１０が再配置指示６２０を出すことである。

【００４９】図１０は、上記論理ディスク再配置要否決定処理部９１０の処理フロー図である。この論理ディスク再配置要否決定処理（９１０）は、ディレクタ１０６が一定周期で各論理ディスク装置２００のアクセス情報５００を検査して行う。ステップ１０００では、アクセス情報５００のアクセス頻度情報５０１を参照し、アクセス頻度が規定値を超え且つ配置されている物理ディスク装置１０５が比較的低速なものである論理ディスク装置（以下、これを第１候補論理ディスク装置という）２００があるか否かをチェックし、該当する論理ディスク装置２００があればステップ１００１へ進み、なければステップ１００５へ進む。

【００５０】ステップ１００１では、前記第１候補論理ディスク装置２００のアクセスパターン情報５０２を参照し、シーケンシャルアクセスの比率が規定値以上であるか否かをチェックし、規定値以上でなければステップ１００２へ進み、規定値以上であればステップ１００４へ進む。

【００５１】ステップ１００２では、前記第１候補論理ディスク装置２００より高速な物理ディスク装置１０５に配置されている論理ディスク装置２００のアクセス頻度情報５０１を参照し、アクセス頻度が規定値以下の論理ディスク装置（以下、これを第２候補論理ディスク装置という）２００があるか否かをチェックし、あればステップ１００３へ進み、なければステップ１００５へ進む。

【００５２】ステップ１００３では、前記第１候補論理ディスク装置２００と前記第２候補論理ディスク装置２００の間で再配置処理（６３０）が必要であると決定し、再配置指示６２０を出す。そして、処理を終了する。

【００５３】ステップ１００４では、前記第１候補論理ディスク装置２００よりシーケンシャル性能の高い物理ディスク装置１０５に配置されている論理ディスク装置２００のアクセスパターン情報５０２を参照し、シーケンシャルアクセスの比率が規定値以下の論理ディスク装置（以下、これを第２候補論理ディスク装置という）２００があるか否かをチェックし、あれば前記ステップ１００３へ進み、なければ前記ステップ１００２へ進む。

【００５４】ステップ１００５では、論理ディスク装置２００の再配置処理（６３０）は不要であると決定する。そして、処理を終了する。

【００５５】以上の第３の実施形態にかかる情報処理システム１および記憶制御装置１０４によれば、アクセス情報５００に基づいて自動的に、アクセス頻度の高い論理ディスク装置をより高速な物理ディスク装置へ再配置することが出来る。また、シーケンシャルアクセスの比率の高い論理ディスク装置をよりシーケンシャルアクセス性能の高い物理ディスク装置へ再配置することが出来る。従って、アクセス性能を向上することが出来る。

【００５６】－第４の実施形態－

上記第１～第３の実施形態を変形して、アクセス情報５００に代えて又は加えて、論理ディスク装置２００に要求される信頼性を再配置処理要否決定の指標に用いてもよい。信頼性を指標に用いれば、論理ディスク装置２００上のデータの信頼性を向上させることが出来る。

【００５７】

【発明の効果】本発明の記憶制御装置によれば、シーケンシャルアクセスの場合やランダムアクセルでヒット率が低い場合でも、アクセス性能を向上することが出来る。また、本発明の記憶制御装置によれば、データの信頼性を向上することが出来る。

【図面の簡単な説明】

【図１】本発明の第１の実施形態にかかる記憶制御装置を含む情報処理システムのブロック図である。

【図２】論理ディスク装置と物理ディスク装置との対応関係の説明図である。

【図３】論理物理対応情報の構成例示図である。

【図４】論理ディスク情報の構成例示図である。

【図５】アクセス情報の構成例示図である。

【図６】本発明の第１の実施形態における記憶制御装置の動作を示すブロック図である。

【図７】論理ディスク装置再配置処理部の処理フロー図である。

【図８】物理ディスク装置アクセス位置算出処理部の処理フロー図である。

【図９】本発明の第３の実施形態における記憶制御装置の動作を示すブロック図である。

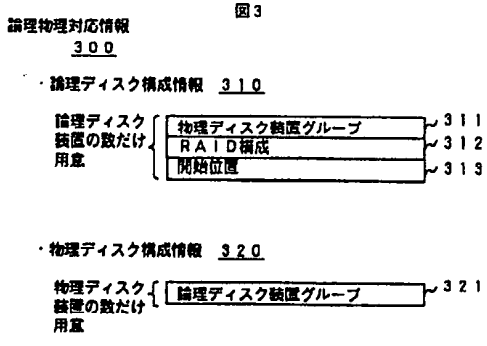
【図１０】論理ディスク装置再配置要否決定処理部の処理フロー図である。

【符号の説明】

１　…情報処理システム
１００…データ処理装置
１０１…ＣＰＵ
１０２…主記憶
１０３…チャネル
１０４…記憶制御装置
１０５…物理ディスク装置
１０６…ディレクタ
１０７…キャッシュメモリ
１０８…キャッシュディレクトリ
１０９…不揮発性メモリ
１１０…不揮発性メモリ管理情報
１１１…ＳＶＰ
２００…論理ディスク装置
２０１…データ
２０２…データ格納位置
３００…論理物理対応情報
４００…論理ディスク情報
５００…アクセス情報
６００…ＣＰＵからの指示
６１０…物理ディスク装置上のアクセス位置算出処理部
６２０…指示情報
６３０…論理ディスク装置再配置処理部
９１０…論理ディスク再配置要否決定処理部

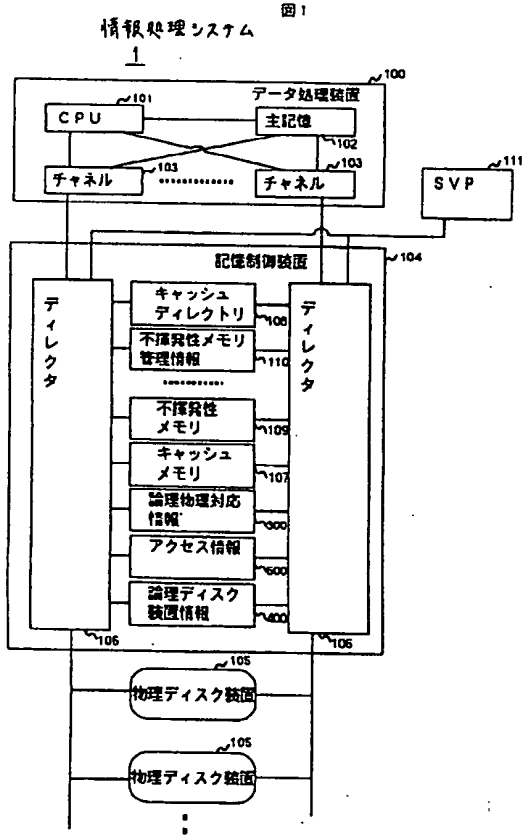
【図３】

図３



【図1】

図1



【図4】

図4

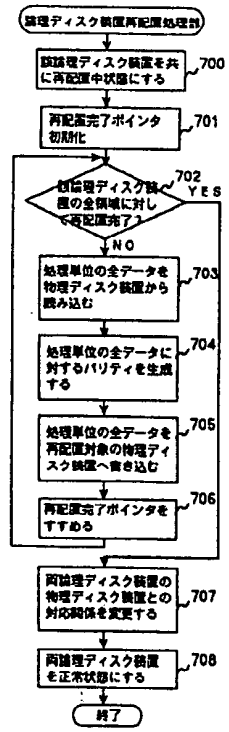
論理ディスク情報

400

論理ディスク装置の数だけ用意	論理ディスク状態	401
	再配置完了ポインタ	402

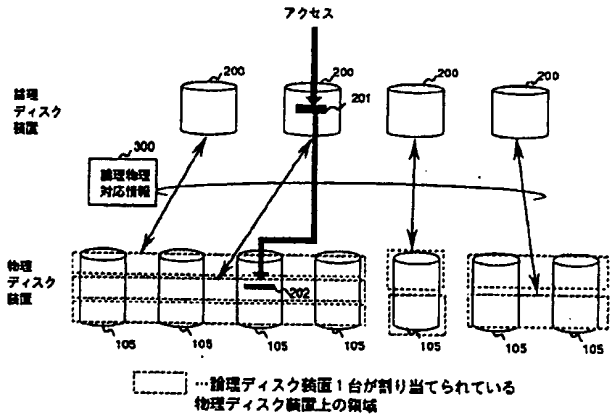
【図7】

図7



【図2】

図2



【図5】

図5

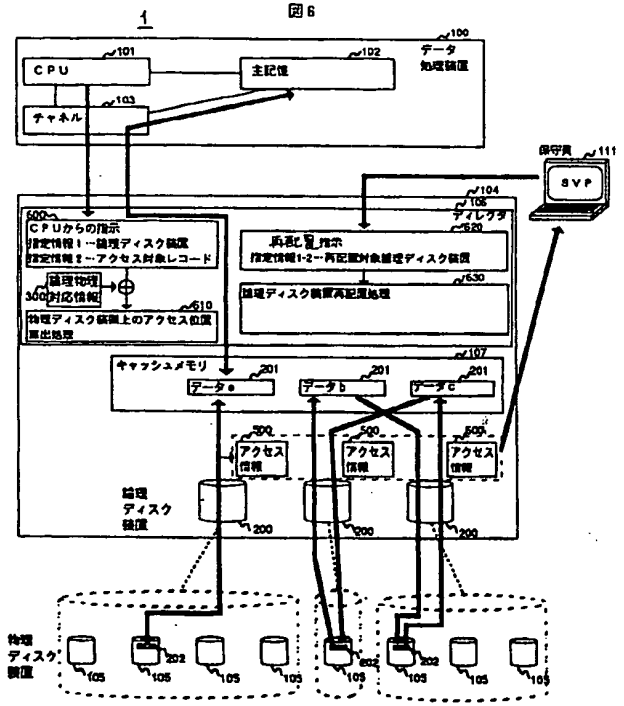
アクセス情報

500

論理ディスク装置の数だけ用意	アクセス頻度情報	501
	アクセスパターン情報	502

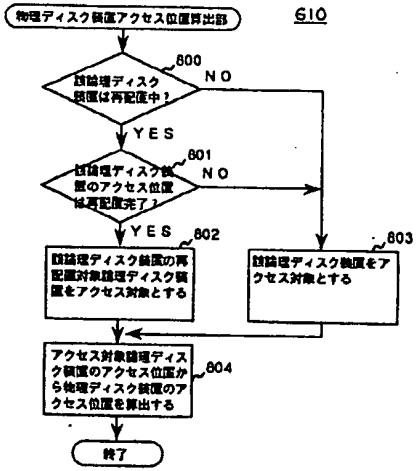
【図6】

図6


1
CPU
主記憶
チャネル
データ処理装置
保守員
SVP
ディレクタ
CPUからの指示
論理物理対応情報
物理ディスク装置上のアクセス位置算出処理
再配置指示
論理ディスク装置再配置処理
キャッシュメモリ
データa
データb
データc
アクセス情報
論理ディスク装置
物理ディスク装置

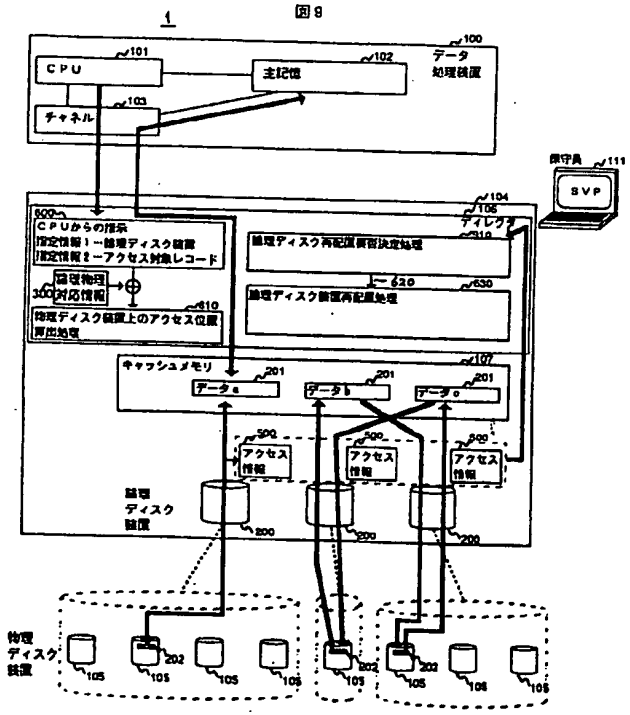
【図8】

図8


物理ディスク装置アクセス位置算出部
610
該論理ディスク装置は再配置中？
NO
YES
該論理ディスク装置のアクセス位置は再配置完了？
NO
YES
該論理ディスク装置の再配置対象論理ディスク装置をアクセス対象とする
該論理ディスク装置をアクセス対象とする
アクセス対象論理ディスク装置のアクセス位置から物理ディスク装置のアクセス位置を算出する
終了

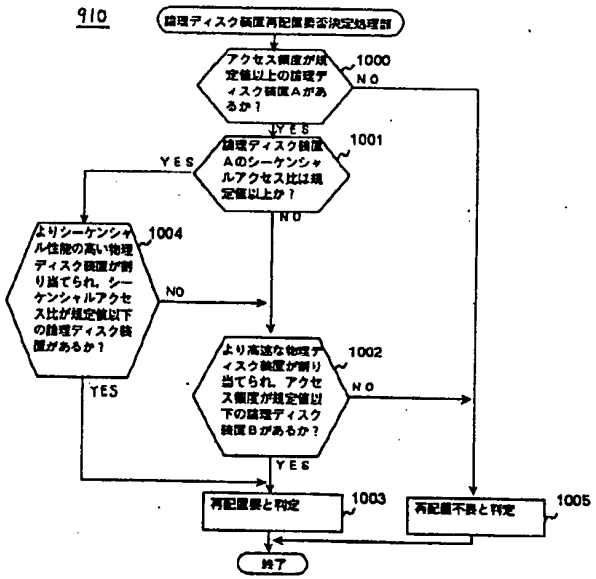
【図9】

図9


1
CPU
101
主記憶
102
データ処理装置
100
チャネル
103
保守員
111
SVP
104
105
ディレクタ
500
CPUからの指示
指定情報1：論理ディスク装置
指定情報2：アクセス対象レコード
論理物理対応情報
300
610
物理ディスク装置上のアクセス位置算出処理
論理ディスク再配置要否決定処理
310
620
論理ディスク装置再配置処理
630
キャッシュメモリ
107
データa
データb
データc
201
アクセス情報
500
論理ディスク装置
200
物理ディスク装置
105
202

【図10】

図10


910
論理ディスク装置再配置要否決定処理部
アクセス頻度が規定値以上の論理ディスク装置Aがあるか？
1000
NO
YES
論理ディスク装置Aのシーケンシャルアクセス比は規定値以上か？
1001
YES
NO
よりシーケンシャル性能の高い物理ディスク装置が割り当てられ、シーケンシャルアクセス比が規定値以下の論理ディスク装置があるか？
1004
NO
YES
より高速な物理ディスク装置が割り当てられ、アクセス頻度が規定値以下の論理ディスク装置Bがあるか？
1002
NO
YES
再配置要と判定
1003
再配置不要と判定
1005
終了

- 動的な領域管理の回避
- ソートのチューニング
- チェックポイント・アクティビティのチューニング
- LGWR および DBWR の I/O のチューニング
- バックアップおよびリストア操作のチューニング
- 大規模プールの構成

I/O の分散によるディスク競合の低減

この項では、ディスク競合を低減する方法を説明します。

- ディスク競合とは
- データ・ファイルと REDO ログ・ファイルの分離
- 表データのストライプ化
- 表と索引の分離
- Oracle と関係のないディスク I/O の低減

ディスク競合とは

複数のプロセスが同時に同じディスクにアクセスしようとするとき、ディスク競合が発生します。多くのディスクには、アクセス数と 1 秒あたりに転送できるデータ量の両方について制限があります。これらの制限に達すると、ディスクにアクセスするためにプロセスを待機させることが必要になります。

通常は、V$FILESTAT ビューの統計およびオペレーティング・システムの機能を検討してください。ディスク性能の限界を判断するために、ハードウェアのマニュアルを調べてください。最大性能やそれに近い性能で作動しているディスクはディスク競合の候補です。たとえば、VMS や UNIX オペレーティング・システム上の一部のディスクでは、1 秒あたり 60 以上の I/O は過剰となる場合があります。

また、db file sequential read、db file scattered read、db file single write、db file parallel write のイベントについて、V$SESSION_EVENT を見なおしてください。これらはすべて、データ・ファイル・ヘッダー、コントローラあるいはデータ・ファイルに対して実行される I/O に対応したイベントです。これらの待機イベントのいずれかが高い平均時間を示している場合は、sar または iostat を使用して I/O 競合を調査してください。次に、デバイスでのビジー待機を探します。ファイル統計を検討して、どのファイルが高い I/O と関連しているかを判定します。

負荷が過剰になっているディスクに対するアクティビティを削減するには、そのディスク上にあるアクセス頻度の激しい 1 つ以上のファイルを、それほど負荷のないディスクに移動します。すべてのディスクの I/O 量がだいたい同じになるまで、各ディスクにこの原則を適用してください。これは、I/O 分散と呼ばれます。

データ・ファイルと REDO ログ・ファイルの分離

Oracle プロセスは、絶えずデータ・ファイルと REDO ログ・ファイルにアクセスします。これらのファイルが同じディスク上に存在している場合、ディスク競合が発生する可能性があります。各データ・ファイルを別々のディスク上に配置してください。そうすると、複数のプロセスがディスク競合せずに同時に異なるファイルにアクセスできます。

REDO ログ・ファイルの各セットは、他のアクティビティがない別々のディスクに配置してください。REDO ログ・ファイルは、トランザクションがコミットされるとき、ログ・ライター・プロセス（LGWR）によって書き込まれます。REDO ログ・ファイル内の情報は順次書き込まれます。同じディスクに対する同時実行のアクティビティが存在しない場合、この順次書込みはさらに高速で行われる可能性があります。REDO ログ・ファイルに別々の専用ディスクを割り当てると、さらにチューニングしなくても通常は LGWR が円滑に実行されます。LGWR に関連するパフォーマンス上のボトルネックはめったにありません。

> **関連項目：** LGWR のチューニングの詳細は、21-14 ページの「REDO ログ・バッファ・ラッチの競合の検出」を参照してください。

データ・ファイル専用ディスクを用意することと REDO ログ・ファイルをミラー化することは、重要な安全対策です。これらの手順を実行することによって、データファイルと REDO ログ・ファイルの両方を単一のディスク障害で失う可能性がないことが保証されます。REDO ログ・ファイルのミラー化によって、REDO ログファイルを単一のディスク障害で失う可能性はないことが保証されます。

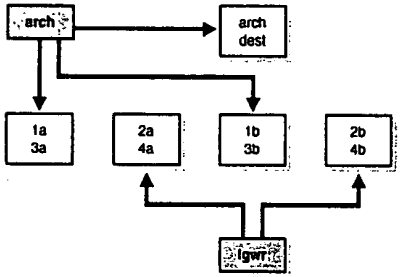
アーカイバ・プロセスと LGWR（マルチ・メンバ・グループを使用している場合）間の I/O 競合を防ぐには、アーカイバの読込みと LGWR の書込みが別個に実行されることを確認してください。たとえば、システムに 2 つのメンバーを持つグループが 4 つある場合、次のシナリオを使用してディスク・アクセスを分離してください。

4 グループ x 2 メンバー = 8 ログ・ファイルで、ラベルは 1a、1b、2a、2b、3a、3b、4a、4b。

それには、最低でも 4 つのディスクと、アーカイブ・ファイル用に 1 つのディスクが必要です。

図 20-2 は、競合を最小限にするために、ディスク間で REDO メンバーを分散する方法を示しています。

図 20-2 ディスク間での REDO メンバーの分散



この例では、LGWR はログ・グループ 1（メンバー 1a と 1b）を切替えて外し、現在はログ・グループ 2（2a と 2b）に書込みを行っています。同時に、アーカイブ・プロセスは、グループ 1 から読込みをして、アーカイブ宛先に書込みを行っています。REDO ログ・ファイルがどのようにして競合から分離されているかに注意してください。

注意： REDO ログ・ファイルをミラー化する、すなわち各 REDO ログ・ファイルの複数のコピーを別々のディスク上に保持することで、LGWR が大幅に遅くはなりません。LGWR は、各ディスクに対して並列して書込みを行い、並列書込みの各部が完了するまで待機します。単一のディスク書込みを実行するために必要な時間は変動することがあるため、コピーの数が増えると、並列書込みでの単一のディスク書込みにかかる時間が平均よりも長くなる傾向が増します。ただし、並列書込みが、最も長い単一のディスク書込みよりも長くなることはありません。また、並列した書込みに関連するオーバーヘッドがオペレーティング・システムで多少発生する場合があります。

表データのストライプ化

ストライプ化、すなわち大きな表のデータを別々のディスク上の別々のデータ・ファイルに分散させることも、競合の低減に役立ちます。

関連項目： この方針については、20-21 ページの「ディスクのストライプ化」で詳しく説明します。

表と索引の分離

頻繁に使用される表は、索引と分離する必要がありません。一連のトランザクション中は、索引が最初に読み込まれてから表が読み込まれます。これらの I/O は順次に発生するので、

表と索引を同じディスク上に格納しても競合は発生しません。ただし、非常に高度な OLTP システムでは、索引と表の分離が必要な場合があります。

索引と表を個別の表領域に分離して、ディスク・ヘッドの移動を最小限に留め、I/O をパラレル化します。すると、1 つのディスク・ヘッドが索引データ上に、残りのヘッドが表データ上にあるので、両読込みともに高速になります。

同時にアクセスされるオブジェクトを分離するという考え方は索引にも当てはまります。たとえば、SQL 文が 2 つの索引を同時に使用する場合、別々のディスクに索引があればパフォーマンスは改善されます。

また、同じディスクに頻繁にアクセスされる複数の表を配置することは避けてください。これを行うには、アプリケーション・アクセス・パターンを熟知している必要があります。

パーティション表とパーティション索引を使用すると、データ・ウェアハウスの操作のパフォーマンスを改善できます。大きな表あるいは索引を異なる表領域に常駐する複数の物理セグメントに分割します。大きなオブジェクト・データ・タイプのある表もすべて個別の表領域に配置してください。

Oracle と関係のないディスク I/O の低減

可能であれば、データベース・ファイルを含むディスクについて、Oracle と関連のない I/O を取り除いてください。この処置は、REDO ログ・ファイルへのアクセスを最適化する上で特に有効です。これによってディスク競合が減少するだけでなく、動的パフォーマンス表 `V$FILESTAT` を使用して、そのようなディスク上のアクティビティをすべて監視することもできます。

ディスクのストライプ化

この項では、次のトピックについて説明します。

- ストライプ化の目的
- I/O のバランス化とストライプ化
- ディスクを手動でストライプ化する方法
- オペレーティング・システム・ソフトウェアでディスクをストライプ化する方法
- ストライプ化と RAID

ストライプ化の目的

ストライプ化によって、大きな表のデータが小さな部分に分割され、これらの部分が別々のディスク上の別々のデータ・ファイルに格納されます。これによって、複数のプロセスがディスク競合なしで表の異なる部分に同時にアクセスできます。ストライプ化は、数多くの行を持つ表へのランダム・アクセスを最適化する上で特に有効です。ストライプ化は、手動で実行する（以降で説明します）ことも、オペレーティング・システムのストライプ化ユーティリティを使用して実行することもできます。

I/O のバランス化とストライプ化

これまで、ベンチマークのチューニング担当者は、使用可能な各デバイス上で I/O の負荷のバランスを均一にすることを懸命に試行していました。現在は、オペレーティング・システムによって、頻繁に使用されるコンテナ・ファイルを多数の物理デバイスにストライプ化する機能が提供されています。ただし、このような手法は、負荷の再分散によってなんらかの形態のキューが排除または削減される場合にのみ有効です。

ドライブでの高ビジー率とともに待機サービス時間が存在する場合には、I/O の分散が必要なことがあります。多数の物理ドライブを使用可能な場合は、2 つの専用ドライブで REDO ログをとることを検討してください（2 つである理由は、REDO ログはオペレーティング・システムまたは Oracle REDO ログ・グループ機能を使用して常にミラー化する必要があるからです）。REDO ログはシリアルに書き込まれるので、REDO ログ・アクティビティ専用のドライブでは通常はヘッドの移動はわずかです。このため、ログ書込みのスピードが大幅に向上します。

アーカイブする場合は、LGWR および ARCH が同じ読込み / 書込みヘッドを競合しないように、別のディスクを使用することが効果的です。これは、ログを代替ドライブに配置することで行います。

ミラー化は、I/O ボトルネックの原因となる可能性もあります。各ミラーへの書込みプロセスは、通常は並列して行われるので、ボトルネックの原因にはなりません。ただし、各ミラーが別々にストライプ化されている場合は、低速のミラー・メンバーが終了するまで I/O は完了しません。I/O の問題を回避するためには、対象データベース（つまりコピー）で元データベースと同数のディスクを使用してストライプ化を行ってください。

たとえば、8 個のディスクに 160KB のデータをストライプ化し、データが 1 つのディスクにしかミラー化されていない場合は、データが 8 個のディスク上でどれだけ高速に処理されるかにかかわらず、160KB がミラー・ディスクに書き込まれるまで I/O は完了しません。したがって、データベースへの書込みには 20.48 ミリ秒しかかかりませんが、ミラーへの書込みには 137 ミリ秒を要します。

ディスクを手動でストライプ化する方法

ディスクを手動でストライプ化するには、オブジェクトの記憶域要件を I/O 要件と関連付ける必要があります。

1. 最初に、次の項目を調べて、オブジェクトのディスク記憶域要件を評価します。
   - オブジェクトのサイズ
   - ディスクのサイズ

   たとえば、オブジェクトが 5GB の Oracle 記憶領域を必要とする場合は、それを収容する 1 つの 5GB のディスクまたは 2 つの 4GB のディスクが必要です。一方、システムが 1GB または 2GB のディスクで構成されている場合は、オブジェクトはそれぞれ 5 個または 3 個のディスクを必要とすることがあります。
2. 20-3 ページの「I/O 要件の分析」で説明したアプリケーションの I/O 要件とこれを比較してください。記憶域要件と I/O 要件の大きい方をとる必要があります。

たとえば、記憶域要件が 5 つのディスク（それぞれ 1GB）であり、I/O 要件が 2 つのディスクである場合は、アプリケーションは大きい方の値である 5 ディスクを必要とします。

3. CREATE TABLESPACE 文で表領域を作成します。DATAFILE 句にデータ・ファイルを指定します。各ファイルは異なるディスク上に作成してください。次に例を示します。

```
CREATE TABLESPACE stripedtabspace
   DATAFILE 'file_on_disk_1' SIZE 1GB,
      'file_on_disk_2' SIZE 1GB,
      'file_on_disk_3' SIZE 1GB,
      'file_on_disk_4' SIZE 1GB,
      'file_on_disk_5' SIZE 1GB;
```

4. CREATE TABLE 文で表を作成します。TABLESPACE 句に新たに作成した表領域を指定します。

   また、STORAGE 句に表のエクステントのサイズを指定します。別々のデータ・ファイルに各エクステントを格納します。表のエクステントは、オーバーヘッドを考慮して表領域内のデータ・ファイルより少し小さくしてください。たとえば、1GB（1024MB）のデータ・ファイルを準備するときは、表エクステントを 1023MB に設定できます。次に例を示します。

```
CREATE TABLE stripedtab (
   col_1  NUMBER(2),
   col_2  VARCHAR2(10) )
   TABLESPACE stripedtabspace
   STORAGE ( INITIAL 1023MB  NEXT 1023MB
      MINEXTENTS 5  PCTINCREASE 0 );
```

   （あるいは、DATAFILE *'datafile'* SIZE *'size'* 句を指定した ALTER TABLE ALLOCATE EXTENT 文を入力することで表をストライプ化することもできます。）

これらのステップによって、表 STRIPEDTAB が作成されます。STRIPEDTAB には、それぞれサイズが 1023MB の初期エクステントが 5 つあります。各エクステントは、CREATE TABLESPACE 文の DATAFILE 句に指定されたデータ・ファイルの 1 つを取り上げます。これらのファイルはすべて別々のディスク上に存在します。MINEXTENTS が 5 なので、これら 5 つのエクステントはすべて即時に割り当てられます。

> **関連項目：** MINEXTENTS および他の記憶域パラメータの詳細は、『Oracle8*i* SQL リファレンス』参照してください。

オペレーティング・システム・ソフトウェアでディスクをストライプ化する方法

手動でディスクをストライプ化する方法のかわりとして、LVM（論理ボリューム・マネージャ）などのオペレーティング・システム・ユーティリティやサード・パーティ・ツールを

使用したり、ハードウェア・ベースのストライプ化機能を使用して、ディスクをストライプ化します。

ユーティリティあるいはハードウェア・ベースのストライプ機構を使用する場合に考慮する主要要因は、ストライプ・サイズ、(ストライプ幅を定義する) ストライプ対象のディスク数および同時実行性のレベル (あるいは I/O アクティビティのレベル) です。これらの要因は、Oracle ブロック・サイズとデータベース・アクセス方法の影響を受けます。

表 20-14 最小ストライプ・サイズ

ディスク・アクセス	最小ストライプ・サイズ
ランダム読込みおよび書込み	最小ストライプ・サイズは、Oracle ブロック・サイズの 2 倍です。
順次読込み	最小ストライプ・サイズは、DB_FILE_MULTIBLOCK_READ_COUNT の値の 2 倍です。

表 20-15 一般ストライプ・サイズ

同時実行性	I/O サイズ	一般ストライプ・サイズ
低	小	k * DB_BLOCK_SIZE
低	大	k * DB_BLOCK_SIZE
高	小	k * DB_BLOCK_SIZE
高	大	k * DB_BLOCK_SIZE * DB_FILE_MULTI_BLOCK_READ_COUNT

ここで、k = 2,3,4... です。

ストライプ化では、データへの統一的なアクセスが前提とされています。ストライプ・サイズが大きすぎる場合は、1 つまたは少数のディスクでホット・スポットが発生する可能性があります。これは、ストライプ・サイズを小さくし、データをより多くのディスクに分散することで回避できます。

固定サイズの 100 行が 5 つのディスクに均一に分散され、各ディスクが 20 の順次行を含んでいる例を考えます。アプリケーションが行 35 ～ 55 へのアクセスのみを必要とする場合は、2 つのディスクのみですべての I/O をする必要があります。同時実行性が高い場合には、システムは目的のパフォーマンス・レベルを達成できない場合があります。

この問題は、行 35 ～ 55 をより多くのディスクに分散することで解決できます。現行の例では、1 ブロックあたりに 2 行とすれば、行 35 と 36 が同じディスク上に存在し、行 37 と行 38 は別のディスクに存在することになります。このアプローチをとると、データはすべてのディスクに分散され、I/O スループットが改善されます。

ストライプ化と RAID

RAID (Redundant arrays of inexpensive disks) 構成では、データの信頼性が改善されます。ただし、I/O パフォーマンスは実装されている RAID 構成によって異なります。

以下に、最も広く使用されている RAID 構成を示します。

- RAID 1: 信頼性および読込み率が向上します。ただし、書込みは不経済になります。
- RAID 0+1: 信頼性が向上し、RAID 1 よりも読込みと書込みのパフォーマンスが向上します。
- RAID 5: 高度の信頼性を提供します。順次読込みを行うと最も多くの利点が引き出せます。書き込みパフォーマンスは RAID 5 ではかなり悪くなります。この構成は、書込み量の多いアプリケーションには推奨できません。

> **注意:** RAID 0 は読込みおよび書込みともに最高のパフォーマンスを提供しますが、冗長性がないため、実際には RAID システムではありません。Oracle では、RAID 0 システムに実働データベース・ファイルを配置しないようお勧めします。

最適なストライプ・サイズは次の 3 項の関数になります。

1. 配列に対する I/O 要求のサイズ
2. 配列に対する I/O 要求の同時実行性
3. ブロック・サイズ境界と一致する物理的なストライプ境界

ストライピングは、配列内の 2 つ以上のディスクへの I/O アクセスを平衡させるには優れたツールです。ただし、次のテクニックを覚えておいてください。

- 同時実行性が高い配列では、単一 I/O 要求が複数の物理 I/O コールに分解されないことを確認する必要があります。そうでなければ、システムで実行される物理 I/O 要求数が何倍にもなり、システム I/O 応答時間が大幅に下がります。
- 同時実行性の低い配列では、単一 I/O が同じディスクに 2 度アクセスすることがないようにする必要があります。同じディスクに 2 度アクセスすると、前述と同じパフォーマンス面でのペナルティが発生します。

動的な領域管理の回避

表やロールバック・セグメントのようなオブジェクトを作成すると、データのために領域がデータベース内に割り当てられます。この領域をセグメントと呼びます。後続のデータベース操作によってデータ容量が増大し、割り当てられた領域を上回るようになると、Oracle はそのセグメントを拡張します。この場合、動的拡張によってパフォーマンスが低下します。

この項では、次のことについて説明します。